

Daniel Thorburn

**LARGE SAMPLE PROPERTIES OF
JACKKNIFE STATISTICS**



Lund 1975

LARGE SAMPLE PROPERTIES OF JACKKNIFE STATISTICS

av

Daniel Thorburn

Fil.kand., Gb

Akademisk avhandling som för avläggande av filosofie doktorsexamen vid matematisk-naturvetenskapliga fakulteten vid universitetet i Lund kommer att offentligen försvaras å matematiska institutionen i Lund, sal MH:A, fredagen den 26 september 1975, kl. 10.15.



ACKNOWLEDGEMENTS

I would like to thank my supervisor, professor Gunnar Blom. His enthusiasm has greatly encouraged me and his constructive criticism has much improved the presentation of the results.

I am also grateful to my colleagues at the department of mathematical statistics for their interest in my work. In particular I would like to mention FK Eva Engvist who has undertaken a thorough scrutiny of my manuscript leading to several improvements and Dr Jan Lanke who has given valuable advice on the language of the thesis.

THE THESIS CONSISTS OF THIS SUMMARY AND THE FOLLOWING THREE PAPERS:

- (1) Thorburn, D.E. Some asymptotic properties of jackknife statistics. Technical report 1975:6, Dept. of Mathematical Statistics, University of Lund, Sweden.
- (2) Thorburn, D.E. On the asymptotic normality of the jackknife. Technical report 1975:7, Dept. of Mathematical Statistics, University of Lund, Sweden.
- (3) Thorburn, D.E. Some different jackknife techniques. Technical report 1975:8, Dept. of Mathematical Statistics, University of Lund, Sweden.

These papers are referred to in the summary as papers 1, 2 and 3.

The sections of these papers are referred to as §§, e.g. § 2.4 means section 4 of paper 2.

1. INTRODUCTION

The jackknife technique is a method to reduce the bias of an estimate. It has been discussed previously by many authors, e.g. by Quenouille (1949, 1956) and Miller (1974). The latter has an excellent bibliography where further references may be found.

Let $g_n, n=1, 2, \dots$ be a sequence of estimates with bias b_n . The bias can often be reduced or, sometimes, eliminated by forming a suitable linear combination; e.g. the expression $ng_n - (n-1)g_{n-1}$ is unbiased if $b_n = 1/n$.

The estimates g_n are generally functions of independent and identically distributed (i.i.d.) random variables (r.v.), $X_1, X_2, X_3, \dots, X_n$, say. It is then natural to require that the combination should be symmetric in the observations. The ordinary jackknife is an example of this:

$$\hat{g}_n = ng_n(X_1, \dots, X_n) - \frac{1}{n} \sum_{i=1}^n (n-1)g_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n).$$

In the i.i.d. situation the jackknife can thus be defined in the following rather vague way:

A jackknife is a linear combination of $g_n, g_{n-1}, \dots$ which is, more or less, symmetric in the observations and formed in order to reduce the bias.

In this summary, results are given only for the i.i.d. case even though exchangeable variables are briefly discussed in papers 1 and 2 (§§ 1.5, 2.3, 2.4, 2.5).

The first two papers contain results only for the ordinary jackknife $\hat{g}_n$ while the last one deals with other types of jackknife. In this summary the results are not given in the same order. Here we bring up a number of properties to discussion and show under what conditions the different jackknives have the property in question.

2. DIFFERENT JACKKNIVES

Several jackknives have been proposed besides $\hat{g}_n$. The latter is a special case of the "complete jackknife based on group size k " (§§ 3.4a, 3.5ab),

$$\hat{g}_n^{(k)} = \frac{1}{k \binom{n}{k}} \sum (ng_n - (n-k)g_{n-k}(x_{i_1}, \dots, x_{i_{n-k}})),$$

where the summation is performed over all subsets of size $n-k$. This statistic behaves differently when the group size is fixed and when it increases with n . The complete jackknife is often difficult to compute and it is mostly of theoretical interest.

More convenient and fairly common is the incomplete jackknife (§§ 3.4b, 3.5a)

$$\hat{g}_{n(k)} = \frac{1}{p} \sum_1^p (ng_n - (n-k)g_{n-k}(x_1, \dots, x_{ik-k}, x_{ik+1}, \dots, x_n))$$

based on $p = n/k$ suitably chosen subsets (p is assumed to be an integer). There often exists a natural grouping, e.g. when the observations are obtained from stratified or cluster sampling (Brillinger 1966). Another incomplete jackknife is (§ 3.5c)

$$\hat{g}_n^v = \frac{1}{qk} \sum_1^k (ng_n - (n-k)g_{n-k}(x_{(i-1)(n-k)+1}, \dots, x_{i(n-k)}))$$

where $q = n/(n-k)$ is an integer.

All the above jackknives remove a bias proportional to $1/n$. A general bias, b_n , is removed by the "generalized jackknife", e.g.

$$\tilde{g}_n = \frac{1}{n} \sum_1^n \frac{b_{n-1}g_n - b_n g_{n-1}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}{b_{n-1} - b_n}$$

or

$$\tilde{g}_n^{(k)} = \frac{1}{\binom{n}{k}} \sum \frac{b_{n-k}g_n - b_n g_{n-k}(x_{i_1}, \dots, x_{i_{n-k}})}{b_{n-k} - b_n}$$

where the summation is over all subsets (§§ 3.2, 3.6).

If the bias consists of two or more linearly independent terms, these terms can be removed by "higher order jackknives" (§ 3.3), i.e. by using one generalized jackknife to remove the first term and then a second one to remove the second term etc.

3. SIMILARITY OF THE JACKKNIFE AND THE ORIGINAL STATISTIC

It is interesting to know when $\hat{g}_n$ and g_n are asymptotically equal. If g_n is the arithmetic mean of n random variables, all jackknives are exactly equal to the original statistic. It is thus natural to try to imitate this situation. Miller (1964) used a smooth function of the arithmetic mean to get something similar to the mean and Arvesen (1969) studied U-statistics.

We use moment conditions for this purpose. The ordinary jackknife is the mean of the pseudo-values,

$$Y_{n,i} = ng_n - (n-1)g_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n).$$

The conditions are, apart from some regularity conditions, (§ 1.2):

$$(1) \quad V(g_n) = \sigma^2/n + o(1/n) \quad \text{as } n \rightarrow \infty.$$

$$(2) \quad E(Y_{n+1} | X_1, \dots, X_n) = r_n \quad \text{where } V(r_n) = o(1/n) \quad \text{as } n \rightarrow \infty.$$

These conditions are sufficient for $V(\hat{g}_n - g_n) = o(1/n)$ as $n \rightarrow \infty$ (§ 1.3). If (1) holds then (2) is almost necessary for this. If conditions (1) and (2) hold, then $\sqrt{n}(f(\hat{g}_n) - f(g_n)) \xrightarrow{d} 0$ and $\sqrt{n}(f_n(\hat{g}_n) - f_n(g_n)) \xrightarrow{d} 0$, where f and f_n are nice functions with bounded second derivatives (§ 1.4).

The same results hold also for generalized and other jackknives based on a fixed group size k (§§ 3.2a, 3.4, 3.6a). On the other hand, if k is a fixed fraction of n , conditions (1) and (2) are not sufficient. Therefore a third condition is introduced:

$$(3) \quad \text{there exist r.v. } Y_i \text{ such that } Y_{n,i} \xrightarrow{ms} Y_i, \quad i = 1, 2, \dots$$

These Y_i , which are independent (§ 2.2), are used for proving that jackknives based on group size k all have the same asymptotic distribution as g_n if (1), (2) and (3) hold (§ 3.5).

When k/n tends to one, these conditions are too strong. If

$$(4) \quad V(g_n) = Cn^{-a} + o(n^{-a}) \quad \text{as } n \rightarrow \infty, \quad C > 0, \quad 0 \leq a \leq 3,$$

then the jackknife and the original statistic have the same asymptotic distribution in mean square: $V(\hat{g}_n^{(k)} - g_n) = V(\hat{g}_n^V - g_n) = o(V(g_n))$ as $n \rightarrow \infty$ (§ 3.5bc).

For higher order jackknives more regularity conditions are needed (§ 3.3). For example the original statistic and the k^{th} order jackknife have the same asymptotic distribution if g_n is a function of the arithmetic mean with $2k+2$ bounded derivatives.

4. ASYMPTOTIC NORMALITY

The variables Y_i defined in (3) are independent (§ 2.2). Their arithmetic mean is thus asymptotically normal. If (1), (2) and (3) hold, the jackknife $\hat{g}_n$ has the same property (§ 2.3). This means that also g_n and the other jackknives are asymptotically normal. Just conditions (1) and (2) are not sufficient for asymptotic normality (§ 2.5).

The mean square convergence in condition (3) can be replaced by convergence in distribution, if there are bounds on the fourth moments of the pseudo-values (§ 2.4). This is shown by means of martingale theory.

5. ESTIMATION OF THE VARIANCE

As a by-product the jackknife often provides a good and simple estimate of the variance; this is true even if there is no bias. For instance, the statistic

$$\frac{1}{n-1} \sum_1^n (Y_{n,i} - \hat{g}_n)^2$$

is a consistent estimate of σ^2 if (1), (2) and (3) hold or if g_n is a smooth function of an estimate for which these conditions hold

(§ 2.2). In paper three similar consistent estimates are given for all situations with a finite group size.

When k/n is fixed the variance estimator corresponding to the incomplete jackknife has an asymptotic χ^2 -distribution if (1), (2) and (3) hold. It is also asymptotically independent of the jackknife itself. Student's t -distribution can thus be used for obtaining confidence regions (§ 3.5ac).

6. A LOWER BOUND ON THE ASYMPTOTIC VARIANCE OF THE JACKKNIFE

Given any sequence of statistics h_n , say, a new sequence g_n may always be found such that $\hat{g}_n = h_n$ (§§ 1.7, 2.5). Such a sequence can be found not only for the ordinary jackknife but for any of the jackknives previously mentioned (§§ 3.2b, 3.6b). If there exists a UMVU-estimate or an estimate optimal in some other way, an original sequence may always be found so that the jackknife has got this property.

The possible improvement is, however, limited. In fact, if (4) holds, then $V(\hat{g}_n)$ cannot be of smaller order than $(3-a)^2 V(g_n)/4$ (§ 1.7). Similar bounds are given for generalized jackknives and jackknives based on increasing group sizes (§§ 3.2b, 3.6b). In particular no type of jackknife has a variance of smaller order than the original statistic if (1) holds.

7. APPLICATIONS

Sample moments obey (1), (2) and (3). In general the jackknife will thus not change the limiting distributions of estimates obtained by the moment method, e.g. a function of the ratio between two sample means (§ 1.6a).

Suppose that g_n is a linear combination of order statistics, $g_n = \sum h((i-1)/(n-1)) X_{i,n}/n$, where h has a bounded second derivative. If $V(X)$ exists, then $\sqrt{n}(\hat{g}_n - g_n) \xrightarrow{ms} 0$ as $n \rightarrow \infty$ (§ 1.6b). If a regularity condition is fulfilled (2) follows. This implies that $\sqrt{n}(f(\hat{g}_n) - f(g_n)) \xrightarrow{d} 0$ if f has a bounded second derivative. On the other hand, if h has a jump (e.g. the trimmed mean), this result is highly improbable (§ 1.6b).

Let g_n be a normalized rank sum, $g_n = \# \{(i,j); X_i \leq Y_j, i,j \leq n\}$. Conditions (1), (2) and (3) hold for this statistic (§ 2.6). All jackknives have thus the same asymptotic normal distribution as g_n . It follows also that the Wilcoxon rank sum is asymptotically normal.

Conditions (1) and (3) are often fairly easy to check but (2) is not always quite so easy. Many statistics can, however, be written as functions of statistics for which (1) and (2) hold. The results of section 1.4 show that the jackknife can be applied without changing the limiting distribution.

The jackknives $\hat{g}_n^{(k)}$ and $\check{g}_n$ have the same limiting distribution as g_n when k/n tends to one if (4) holds. This condition holds for almost all statistics, e.g. for the median of a continuous distribution and for the minimum of a uniform distribution (§ 3.5b).

REFERENCES

- Arvesen, J.N. (1969). Jackknifing U-statistics. *Ann. Math. Statist.* 40, p 2076 - 2100.
- Brillinger, D.R. (1966). The application of the jackknife to the analysis of sample surveys. *Commentary* 8, p 74-80.
- Miller, R.G. (1964). A trustworthy jackknife. *Ann. Math. Statist.* 35, p 1594 - 1605.
- Miller, R.G. (1974). The jackknife - a review. *Biometrika* 61, p 1-15.
- Quenouille, M.H. (1949) Approximate tests of correlation in time-series. *J. R. Statist. Soc. B* 11, p 68 - 84.
- Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika* 43, p 353 - 360.

